Supplementary Notes

-

Discovering Millions of Plankton Genomic Markers from the Atlantic Ocean and the Mediterranean Sea

Majda Arif[1], Jérémy Gauthier[2], Kevin Sugier[1], Daniele Ludicone[3], Olivier Jaillon[1], Patrick Wincker[1], Pierre Peterlongo[2], Mohammed-Amin Madoui[1]

[1] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France

[2] Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes

[3]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121, Naples, Italy

Supplementary Notes S1: SRA accession numbers and sampling coordinates of metagenomic data.

Location	Station	Latitude	Longitude	Fraction size (µm)	Sample accession	Secondary sample accession	Run accession
Mediterranean Sea	4	36.0552	-6.56275	0.8-5	SAMEA2619396	ERS487919	ERR868369
				5-20	SAMEA2619398	ERS487921	ERR1726542
				180-2000	SAMEA2657059	ERS487930	ERR1726827
Mediterranean Sea	7	37.04875	1.9402	0.8-5	SAMEA2591060	ERS477934	ERR315802 ERR315821
				20-180	SAMEA2611380	ERS477943	ERR538183 ERR315827
Mediterranean Sea	8	38.004	3.97775	20-180	SAMEA2730447	ERS488104	ERR1756211
Mediterranean Sea	9	39.12085	5.85945	0.8-5	SAMEA2619534	ERS488122	ERR868407
				180-2000	SAMEA2657011	ERS488134	ERR1726662 ERR1726681
Mediterranean Sea	10	40.64115	2.87725	20-180	SAMEA2730458	ERS488205	ERR1718269
Mediterranean Sea	11	41.1686	2.7996	0.8-2000	SAMEA2619621	ERS488249	ERR1726597
				5-20	SAMEA2619623	ERS488251	ERR1700895
				20-180	SAMEA2656955	ERS488262	ERR1726642 ERR1726791
				180-2000	SAMEA2657003	ERS488257	ERR1726953 ERR1726879
Mediterranean Sea	12	43.3515	7.8994	20-180	SAMEA2730644	ERS488267	ERR1718362
Mediterranean Sea	18	35.7491	14.29475	0.8-5	SAMEA2619675	ERS488338	ERR868393
				20-180	SAMEA2656983	ERS488342	ERR1726956
				5-20	SAMEA2619676	ERS488339	ERR1726555
Mediterranean Sea	20	34.44785	14.9261	180-2000	SAMEA2657025	ERS488420	ERR1726786
				5-20	SAMEA2619725	ERS488412	ERR1700896
Mediterranean Sea	22	39.83935	17.41535	0.8-5	SAMEA2619745	ERS488446	ERR868403
				5-20	SAMEA2619746	ERS488447	ERR1726540 ERR1726929
				20-180	SAMEA2657062	ERS488453	ERR1726613 ERR1726546
				180-2000	SAMEA2731167	ERS488468	ERR1726802 ERR1726810
Mediterranean Sea	23	42.1956	17.71695	0.8-5	SAMEA2591095	ERS477990	ERR538173 ERR318582
				20-180	SAMEA2624979	ERS478009	ERR318594 ERR318583 ERR538171
				180-2000	SAMEA2624982	ERS477997	ERR538185 ERR318600
Mediterranean Sea	24	42.45705	17.94285	20-180	SAMEA2656978	ERS488481	ERR1726675
Mediterranean Sea	25	39.37435	19.39945	0.8-5	SAMEA2619777	ERS488497	ERR868356
				5-20	SAMEA2619778	ERS488498	ERR1726722
				20-180	SAMEA2656999	ERS488503	ERR1726976
				180-2000	SAMEA2657015	ERS488506	ERR1726787
Mediterranean Sea	26	38.4493	20.18135	0.8-20	SAMEA2619798	ERS488534	ERR868402
				20-180	SAMEA2657002	ERS488539	ERR1726673

Supplementary Notes S1: SRA accession numbers and sampling coordinates of metagenomic data (suite).

Oceanographic Region	Station	Latitude	Longitude	Fraction size (µm)	Sample accession	Secondary sample accession	Run accession
Mediterranean Sea	30	33.9191	32.86955	5-20	SAMEA2591118	ERS478027	ERR538178 ERR318589 ERR318602
				20-180	SAMEA2624983	ERS478034	ERR318616 ERR318585 ERR538189
				180-2000	SAMEA2624978	ERS478037	ERR318609 ERR318614 ERR318608 ERR318610 ERR318587 ERR318596
Southern Atlantic	66	-34.92175	17.97445	0.8-5	SAMEA2620939	ERS490134	ERR599261
				5-20	SAMEA2620941	ERS490136	ERR599208
				20-180	SAMEA2657046	ERS490159	ERR599301 ERR1726663
				180-2000	SAMEA2656956	ERS490150	ERR1726689 ERR599277
Southern Atlantic	67	-32.2176	17.70515	0.8-5	SAMEA2620988	ERS490201	ERR599302
				5-20	SAMEA2657014	ERS490220	ERR599295
				20-180	SAMEA2657065	ERS490216	ERR599269
				180-2000	SAMEA2656969	ERS490210	ERR599255
Southern Atlantic	68	-31.0379	4.66455	0.8-5	SAMEA2621029	ERS490281	ERR599257
				5-20	SAMEA2657009	ERS490261	ERR599223
				20-180	SAMEA2657082	ERS490256	ERR599304
				180-2000	SAMEA2657073	ERS490253	ERR599243
Southern Atlantic	70	-20.40335	-3.1884	0.8-5	SAMEA2621082	ERS490343	ERR599305
				5-20	SAMEA2656958	ERS490366	ERR599241 ERR1726969
				20-180	SAMEA2656971	ERS490362	ERR599258 ERR1726899
				180-2000	SAMEA2657034	ERS490356	ERR599313
Southern Atlantic	72	-8.81015	-17.91365	5-20	SAMEA2656989	ERS490467	ERR599204
				20-180	SAMEA2656961	ERS490463	ERR599256
				180-2000	SAMEA2657090	ERS490457	ERR599286
Southern Atlantic	76	-20.97135	-35.25735	5-20	SAMEA2657095	ERS490582	ERR1726626 ERR599332
				20-180	SAMEA2657030	ERS490572	ERR1726937 ERR599326 ERR1726731
				180-2000	SAMEA2657091	ERS490564	ERR1726678 ERR599336

Supplementary Notes S1: SRA accession numbers and sampling coordinates of metagenomic data (suite).

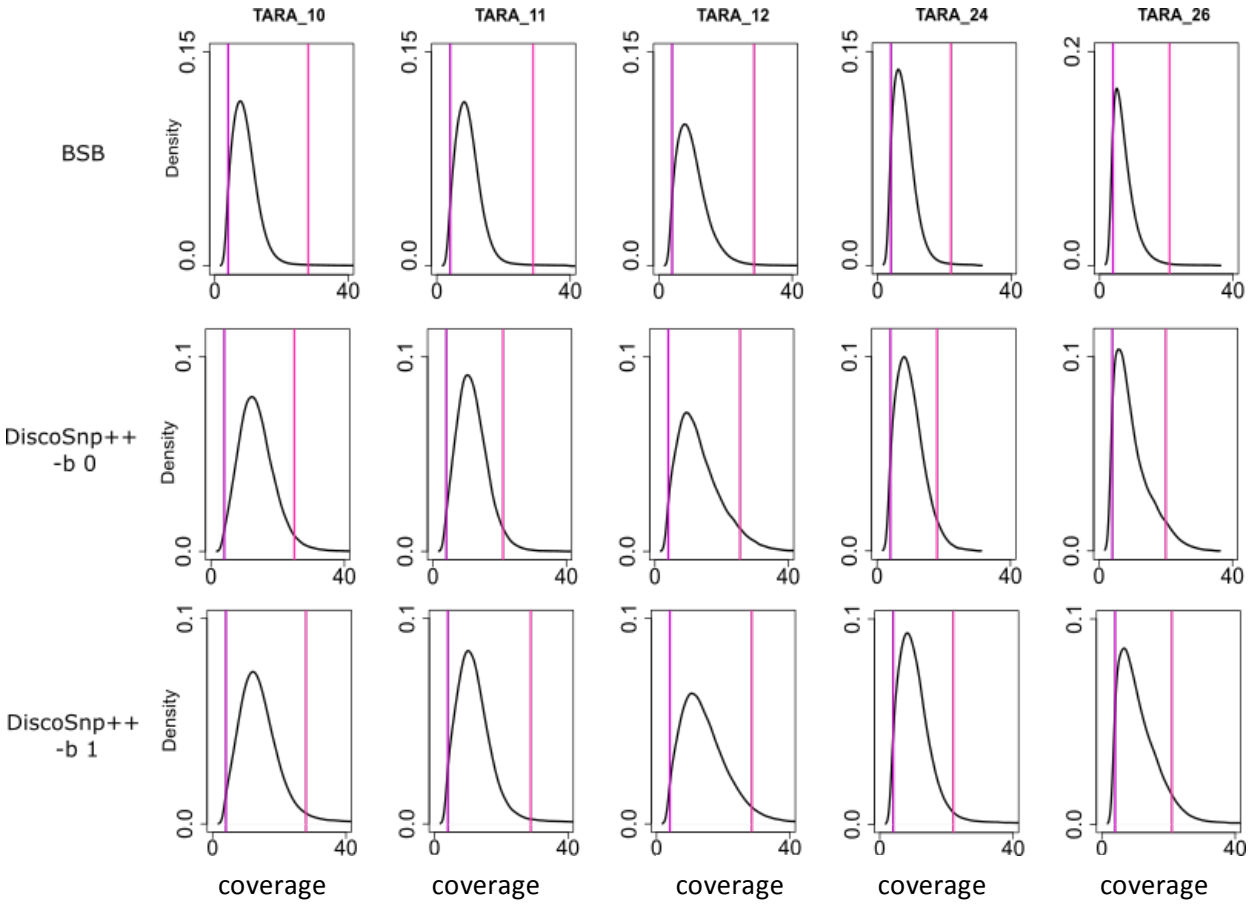
Location	Station	Latitude	Longitude	Fraction size (µm)	Sample accession	Secondary sample accession	Run accession
Southern Atlantic	78	-30.1626	-43.2865	5-20	SAMEA2657029	ERS490687	ERR599250
				20-180	SAMEA2657067	ERS490683	ERR599263
				180-2000	SAMEA2656977	ERS490679	ERR599245
Southern Atlantic	80	-40.65025	-52.1683	0.8-5	SAMEA2621315	ERS490751	ERR868387
				20-180	SAMEA2732299	ERS490768	ERR1726636
				180-2000	SAMEA2732231	ERS490761	ERR1726819
Southern Atlantic	81	-44.5366	-52.43165	0.8-5	SAMEA2621362	ERS490817	ERR868372
				20-180	SAMEA2732891	ERS490871	ERR1726853
Southern Atlantic	82	-47.18845	-58.2792	0.8-5	SAMEA2621412	ERS490896	ERR599298
				5-20	SAMEA2657104	ERS490915	ERR599232
				20-180	SAMEA2657078	ERS490911	ERR599306
				180-2000	SAMEA2656994	ERS490905	ERR599234
Austral Ocean	83	-54.3766	-65.07765	0.8-5	SAMEA2621470	ERS490977	ERR868388
				5-20	SAMEA2621474	ERS490981	ERR1700898
				180-2000	SAMEA2732412	ERS490995	ERR1726928
Austral Ocean	84	-60.29465	-60.5786	0.8-5	SAMEA2621498	ERS491012	ERR599254
				5-20	SAMEA2657096	ERS491032	ERR599224
				20-180	SAMEA2657033	ERS491028	ERR599283
				180-2000	SAMEA2657049	ERS491021	ERR599291
Austral Ocean	85	-62.0874	-49.43105	0.8-5	SAMEA2621522	ERS491057	ERR599335
				20-180	SAMEA2732415	ERS491073	ERR599264
				180-2000	SAMEA2656979	ERS491070	ERR599214
Northern Atlantic	142	25.546	-88.42815	0.8-5	SAMEA2623479	ERS493954	ERR868430
				20-180	SAMEA2730804	ERS493971	ERR1726768
				180-2000	SAMEA2730918	ERS493965	ERR1726944
Northern Atlantic	143	29.6871	-79.60905	5-20	SAMEA2730950	ERS494049	ERR1700893
				20-180	SAMEA2730956	ERS494055	ERR1726737
				180-2000	SAMEA2730657	ERS494045	ERR1726638
Northern Atlantic	144	36.36745	-72.86875	0.8-5	SAMEA2623603	ERS494131	ERR873964
				5-20	SAMEA2731011	ERS494136	ERR1700902
				20-180	SAMEA2730686	ERS494142	ERR1726581
				180-2000	SAMEA2730691	ERS494147	ERR1726654
Northern Atlantic	145	39.2239	-71.0352	0.8-5	SAMEA2623641	ERS494184	ERR868411
				5-20	SAMEA2730594	ERS494199	ERR1726547
				20-180	SAMEA2731035	ERS494195	ERR1726971
				180-2000	SAMEA2731031	ERS494191	ERR1726885
Northern Atlantic	146	34.75265	-71.25665	0.8-5	SAMEA2623685	ERS494248	ERR868351
				5-20	SAMEA2730877	ERS494263	ERR1726548
				20-180	SAMEA2730873	ERS494259	ERR1726757
				180-2000	SAMEA2730869	ERS494255	ERR1726764
Northern Atlantic	147	33.01575	-66.53725	0.8-5	SAMEA2623723	ERS494304	ERR868366
				5-20	SAMEA2732078	ERS494323	ERR1726703
				20-180	SAMEA2732044	ERS494317	ERR1726720
				180-2000	SAMEA2731996	ERS494313	ERR1726877
							ERR1726831

Supplementary Notes S1: SRA accession numbers and sampling coordinates of metagenomic data (suite).

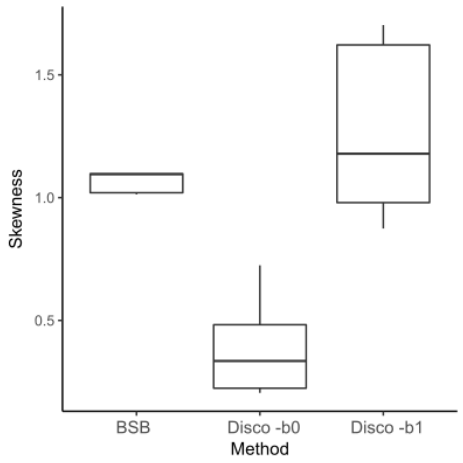
Location	Station	Latitude	Longitude	Fraction size (µm)	Sample accession	Secondary sample accession	Run accession
Northern Atlantic	148	31.7657	-64.17895	5-20	SAMEA2731821	ERS494359	ERR1726534
				20-180	SAMEA2731634	ERS494355	ERR1726625
				180-2000	SAMEA2732099	ERS494350	ERR1726889
Northern Atlantic	149	34.1129	-49.92245	5-20	SAMEA2731851	ERS494422	ERR1726772
				20-180	SAMEA2731641	ERS494418	ERR1726867
				180-2000	SAMEA2731846	ERS494412	ERR1726949
Northern Atlantic	150	35.89685	-37.238	0.8-5	SAMEA2623817	ERS494454	ERR868354
				5-20	SAMEA2732095	ERS494470	ERR1726973
				20-180	SAMEA2731788	ERS494465	ERR1726860
				180-2000	SAMEA2731784	ERS494461	ERR1726859
Northern Atlantic	151	36.15525	-28.9914	0.8-5	SAMEA2623861	ERS494529	ERR868459
				5-20	SAMEA2732085	ERS494545	ERR1726847
				20-180	SAMEA2732056	ERS494540	ERR1726744
				180-2000	SAMEA2732052	ERS494536	ERR1726712
Northern Atlantic	152	43.6849	-16.84495	0.8-5	SAMEA2623901	ERS494594	ERR868445
				5-20	SAMEA2731676	ERS494606	ERR1726861
				20-180	SAMEA2731859	ERS494612	ERR1726635
				180-2000	SAMEA2731672	ERS494602	ERR1726553

Supplementary Notes S2: Methods effect on the biallelic loci coverage. a. Coverage distribution and cut-offs used to select variants. The blue lines correspond to the loci coverage where the variants where detected**.** The purple and pink vertical lines correspond to lower and upper limit values for locus coverage to select a variant. **b.** Effect of the method on the skewness of the biallelic loci. **c.** Effect of the method on the difference between the expected and the observed skewness.

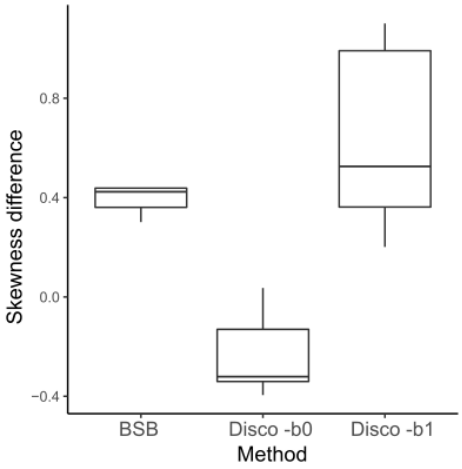
a



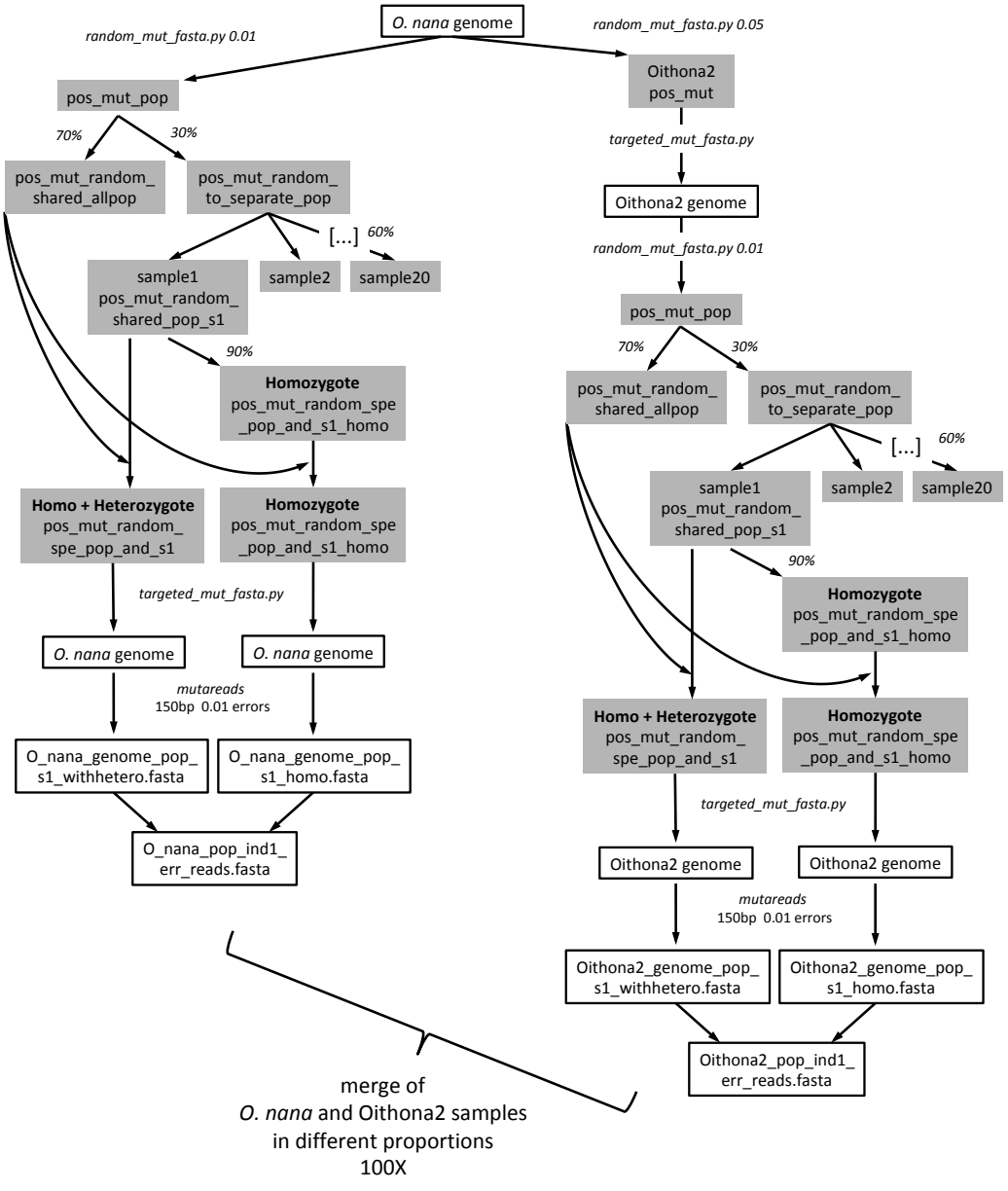
b



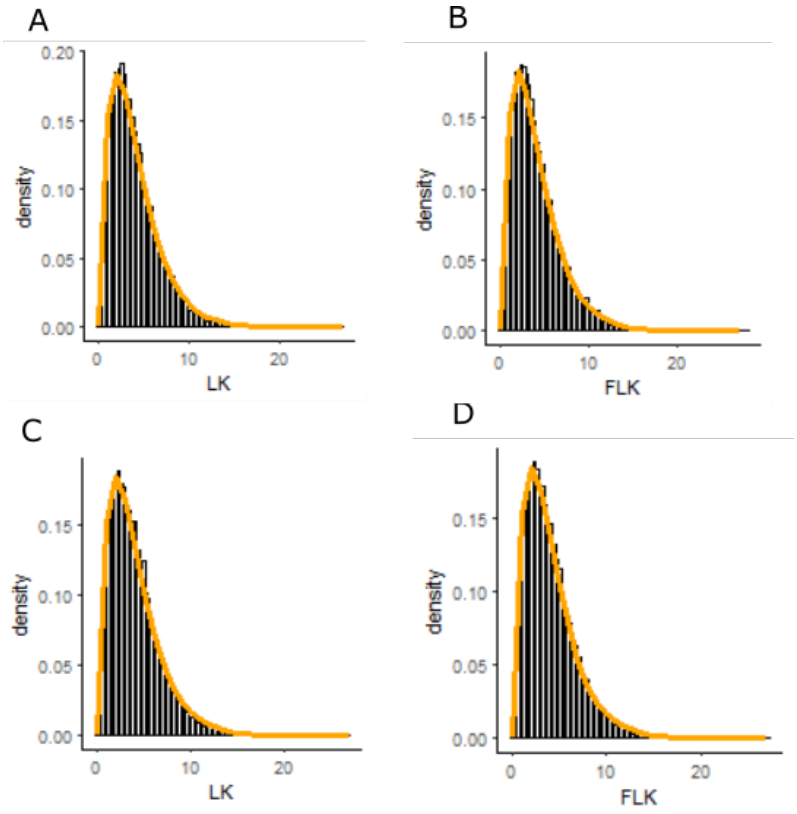
c



Supplementary Notes S3: Variant calling based on simulated data. To produce simulated data, the O. nana genome sequence was first mutated inducing 1% divergence. Seventy percent of these SNPs were shared by all samples to simulate population-specific polymorphism. The other 30% were randomly distributed across 20 individuals including shared polymorphism, i.e. for each sample, 60% of SNPs are randomly picked in the SNPs set. Moreover, in each sample, 10% of the polymorphism was introducing as heterozygote. Then, 150-bp Illumina reads were generated introducing sequencing errors. For Oithona2 samples, O. nana genome was first mutated inducing 5% divergence. Then, this new Oithona2 genome was mutated at 1% to generate 20 individuals. After this first step, SNPs were distributed in a population as for the O. nana individuals. Finally, to simulate metagenomic data, reads from the 40 samples (20 O. nana individuals and 20 Oithona2 individuals), were merged at different proportions from 0% to 100% by 5%, and 30X of reads were samples for each admixture. The code used for the simulation is available at https://github.com/GATB/DiscoSnp/tree/master/scripts/simulations



Supplementary Notes S4: LK and FLK distribution obtained from BSB and *DiscoSnp++* BAFs. The theoretical χ^2 distribution (with df=4) is in orange and the observed LK and F_{ST} are black bars. **a.** LK distribution from the BSB BAFs. **b.** FLK distribution from the BSB BAFs. **c.** LK distribution from the *DiscoSnp++* BAFs. **d.** FLK distribution from the *DiscoSnp++* BAFs.

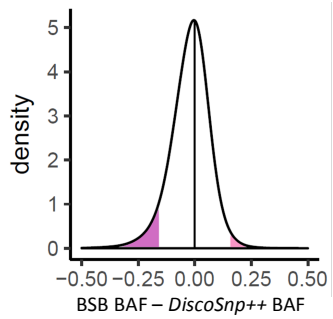


Supplementary Notes S5: Results of the methods comparison simulated data with *O. nana* and another closely related species admixture in different proportions

	BSB		*DiscoSnp++*					
%age of *O. nana*	*O. nana* variants	Oithona2 variants	*O. nana* variants	Oithona2 variants	BSB TP/ DiscoSnp++ TP	BSB FP/ DiscoSnp++ FP	BSB SN	DiscoSnp++ SN
0	2,089	206,657	1,249	131,962	1.67	1.57	0.01	0.01
5	92,206	201,090	1,161	122,565	79.42	1.64	0.46	0.01
10	294,857	193,912	1,119	111,852	263.50	1.73	1.52	0.01
15	466,532	187,260	1,199	101,861	389.10	1.84	2.49	0.01
20	579,023	185,177	1,498	91,491	386.53	2.02	3.13	0.02
25	651,396	188,863	2,220	80,896	293.42	2.33	3.45	0.03
30	697,749	195,782	3,492	71,689	199.81	2.73	3.56	0.05
35	728,083	201,319	5,779	62,001	125.99	3.25	3.62	0.09
40	749,797	202,651	9,531	53,253	78.67	3.81	3.70	0.18
45	764,168	197,488	14,998	44,540	50.95	4.43	3.87	0.34
50	774,396	187,633	22,199	36,515	34.88	5.14	4.13	0.61
55	781,932	173,799	31,169	29,927	25.09	5.81	4.50	1.04
60	787,358	153,802	41,244	23,807	19.09	6.46	5.12	1.73
65	791,509	130,268	52,922	19,060	14.96	6.83	6.08	2.78
70	794,911	106,559	64,785	14,962	12.27	7.12	7.46	4.33
75	797,479	81,286	77,558	11,647	10.28	6.98	9.81	6.66
80	799,705	57,640	89,973	9,172	8.89	6.28	13.87	9.81
85	801,572	36,824	102,022	7,769	7.86	4.74	21.77	13.13
90	803,048	20,043	114,089	7,275	7.04	2.76	40.07	15.68
95	804,434	8,875	125,427	7,736	6.41	1.15	90.64	16.21
100	805,671	4,072	135,111	8,115	5.96	0.50	197.86	16.65

Supplementary Notes S6: Differences in genomic location of variants presenting a high difference in BAF values between the two variant calling methods. a. Distribution of the BAF difference between *DiscoSnp++* and BSB, the purple zone correspond to outlier loci. **b.** Count of variants depending on their BAFs correlation between the two methods. The set A corresponds to the variants that have a similar *DiscoSnp++* and BSB BAF. The set B corresponds to the variants that have a higher BAF value with *DiscoSnp++*. The set C corresponds to the variants that have higher BAF value with BSB. **c.** Barplot of the relative counts with the set D that corresponds to a random distribution of the variants in the *O. nana* genome. The stars indicate significant differences in variants proportion (*p*-value < 0.001, chi-square test).

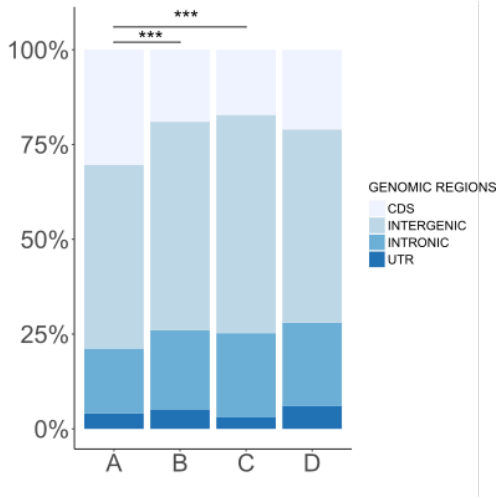
a



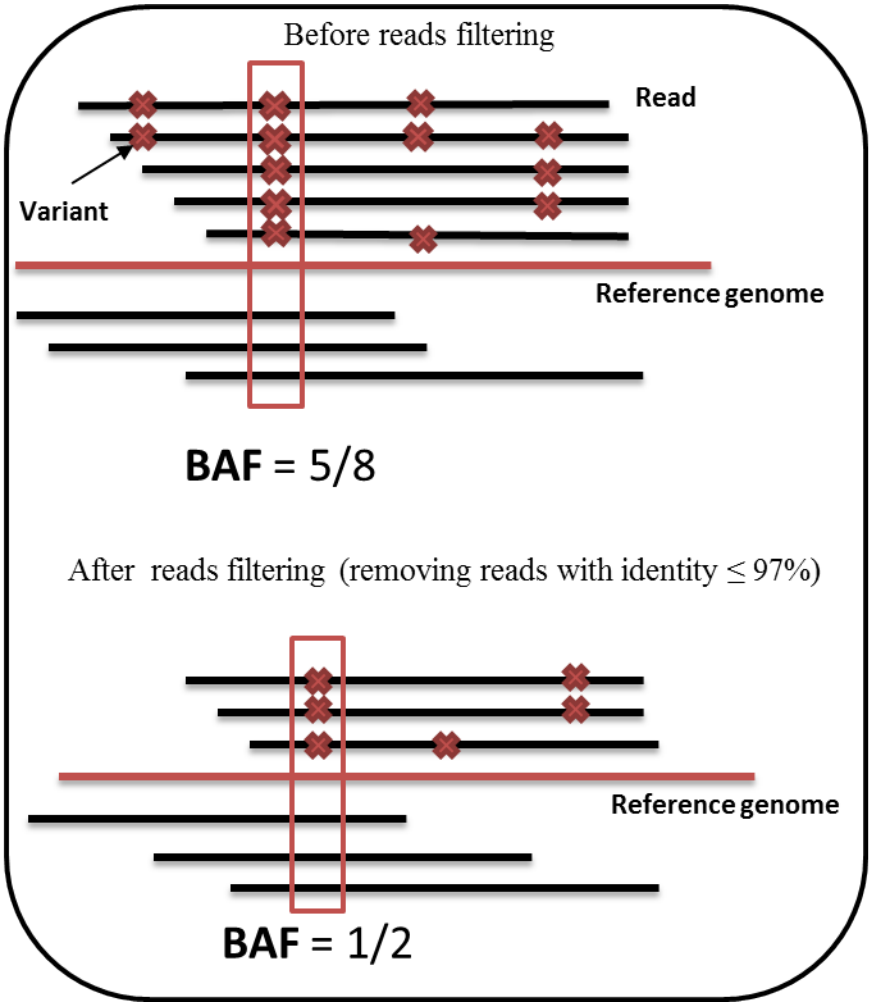
b

Genomic location	Correlated variants (Set A)	Negative outliers (set B)	Positive outliers (Set C)
UTR	2 602	190	26
CDS	17 229	730	130
Intron	10 017	806	169
Intergenic	27 391	2 152	429
Total	**57 239 (92.5%)**	**3 878 (6.3%)**	**754 (1.2%)**

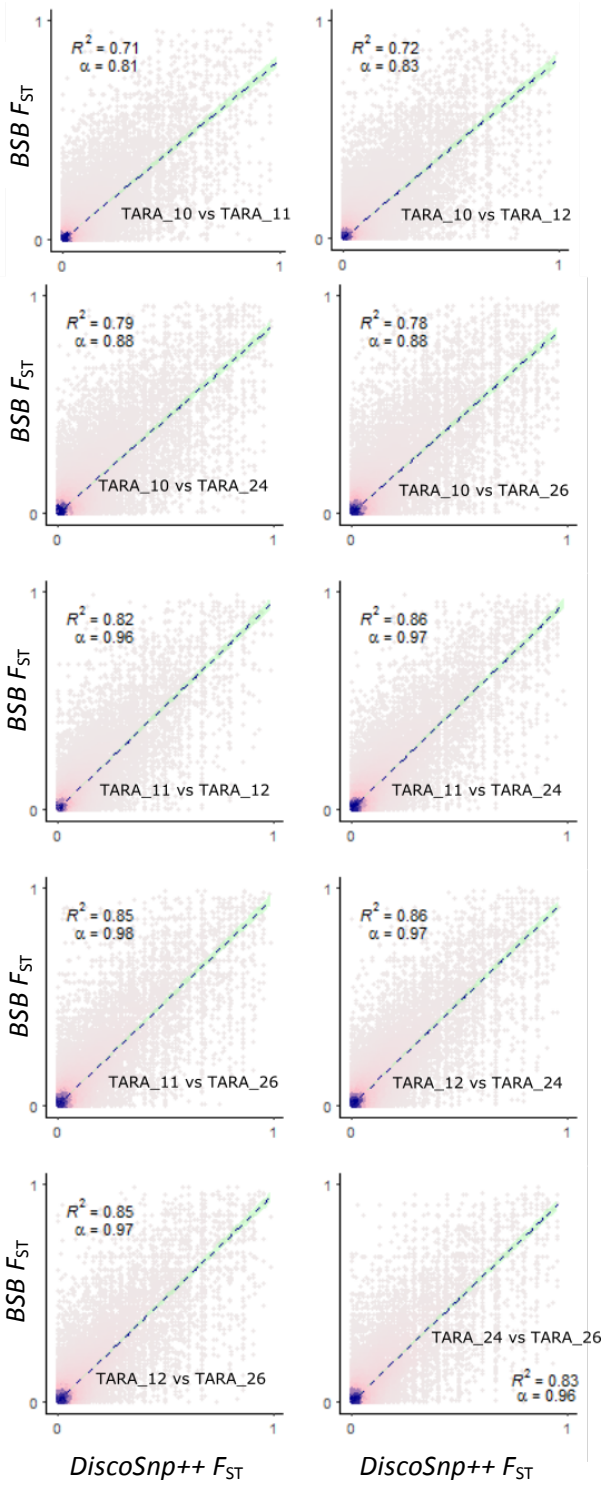
c



Supplementary Notes S7: Effect of the reads filtering based the identity cut-off on the BAF calculation by BSB.



Supplementary Notes S8: Scatter-plot of the pairwise F_{ST} computed from BSB and DiscoSnp using the using BAFs from set1 and set2. The blue dashed line is the linear regression and the green lines are the confidence interval. The x-axis corresponds to *DiscoSnp++* F_{ST} and the y-axis corresponds to BSB F_{ST}.



Supplementary Notes S9: Annotation of loci detected under natural selection.

Scaffold	Position	Ref	Alt	Gene model	Variant	AA modification	Annotation	Lk.b	pval.Lk.b	Lk.d	pval.Lk.d
1	218542	C	T		Upstream gene			2.37E+01	9.37E-05	1.89E+01	8.04E-04
1	467134	C	A		Upstream gene			1.95E+01	6.25E-04	1.91E+01	7.66E-04
2	966381	G	C	GSONAT00000516001	NA		Serine-pyruvate aminotransferase	1.81E+01	1.18E-03	1.72E+01	1.78E-03
8	891235	T	C		NA		None predicted	2.03E+01	4.45E-04	1.93E+01	6.91E-04
8	891330	T	A		NA		None predicted	2.01E+01	4.84E-04	1.92E+01	7.16E-04
9	971651	A	T	GSONAT00015360001	NA		Pickpocket protein like	1.80E+01	1.25E-03	1.71E+01	1.81E-03
10	237360	T	C	GSONAT00002382001	Synonymous	p.Leu2027Leu	Dynein heavy chain	2.06E+01	3.78E-04	1.95E+01	6.18E-04
10	247395	A	G	GSONAT00002383001	Upstream gene		None predicted	2.10E+01	3.20E-04	1.87E+01	8.89E-04
10	247465	G	A	GSONAT00002383001	Synonymous	p.Phe19Phe	None predicted	2.23E+01	1.77E-04	1.90E+01	7.90E-04
10	247594	A	C	GSONAT00002383001	5 prime UTR		None predicted	1.94E+01	6.47E-04	1.86E+01	9.60E-04
10	247669	G	T	GSONAT00002383001	5 prime UTR		None predicted	2.10E+01	3.20E-04	2.00E+01	5.08E-04
15	277847	A	T	GSONAT00003212001	Downstream gene			1.80E+01	1.22E-03	2.00E+01	5.02E-04
23	372498	T	C		Upstream gene			1.80E+01	1.22E-03	1.76E+01	1.47E-03
25	231999	G	A	GSONAT00005206001	5 prime UTR		Fermitin family	1.82E+01	1.15E-03	1.71E+01	1.87E-03
51	153302	C	T	GSONAT00008031001	synonymous	p.Ser351Ser	Secretin-like	2.17E+01	2.26E-04	1.97E+01	5.85E-04
67	27053	T	C		Upstream gene			1.84E+01	1.05E-03	1.60E+01	3.02E-03
68	10871	T	C		Upstream gene			1.75E+01	1.54E-03	1.80E+01	1.25E-03
75	294967	G	A	GSONAT00009904001	Synonymous	p.Cys213Cys	PAN PAN/Appel domain	1.98E+01	5.58E-04	1.74E+01	1.61E-03
75	295105	G	A	GSONAT00009904001	Synonymous	p.Phe167Phe	PAN PAN/Appel domain	2.12E+01	2.93E-04	2.00E+01	5.02E-04
75	295475	T	C	GSONAT00009904001	Missense	p.Thr64Ala	PAN PAN/Appel domain	2.37E+01	9.37E-05	2.23E+01	1.72E-04
75	304202	C	T	GSONAT00009907001	Synonymous	p.Pro11Pro	ARL14 effector protein	1.96E+01	6.01E-04	1.97E+01	5.85E-04
75	304522	T	C	GSONAT00009907001	5 prime UTR		ARL14 effector protein	1.74E+01	1.59E-03	1.73E+01	1.71E-03
75	304956	T	G	GSONAT00009908001	Synonymous	p.Ala196Ala	None predicted	2.17E+01	2.26E-04	2.01E+01	4.77E-04
76	166168	T	C	GSONAT00009935001	Synonymous	p.Thr536Thr	Kelch-type beta propeller domain	1.85E+01	9.80E-04	1.65E+01	2.38E-03
76	166248	A	G	GSONAT00009935001	Missense	p.Glu563Gly	Kelch-type beta propeller domain	1.72E+01	1.74E-03	1.60E+01	3.08E-03
76	166428	A	G	GSONAT00009935001	Missense	p.Lys623Arg	Kelch-type beta propeller domain	2.04E+01	4.22E-04	2.02E+01	4.46E-04
86	141392	C	T	GSONAT00010501001	Intron		RasGRF2	1.92E+01	7.09E-04	1.84E+01	1.01E-03
86	141429	A	G	GSONAT00010501001	Intron		RasGRF2	2.22E+01	1.86E-04	2.00E+01	5.08E-04
86	142467	T	A	GSONAT00010501001	Intron		RasGRF2	1.81E+01	1.17E-03	1.86E+01	9.28E-04
88	109848	T	G		Upstream gene			1.77E+01	1.44E-03	1.75E+01	1.57E-03
94	105665	G	A	GSONAT00015420001	NA		TNF-like domain superfamily	2.37E+01	9.37E-05	2.23E+01	1.72E-04
94	109199	C	T		Upstream gene			2.37E+01	9.37E-05	2.23E+01	1.72E-04
94	113922	T	C	GSONAT00010837001	Missense	p.Val11Ala	Sugar transporter-like	2.04E+01	4.22E-04	1.62E+01	2.81E-03
102	143819	A	T	GSONAT00011159001	NA		Arylsulfatase	2.20E+01	1.96E-04	2.03E+01	4.43E-04
102	159878	G	T	GSONAT00011161001	Missense	p.Gln382Lys	FMRFamide receptor	1.89E+01	8.25E-04	1.74E+01	1.64E-03
Scaffold	Position	Ref	Alt	Gene model	Variant	AA modification	Annotation	Lk.b	pval.Lk.b	Lk.d	pval.Lk.d
103	208056	C	T	GSONAT00011184001	Synonymous	p.Leu412Leu	None predicted	1.81E+01	1.18E-03	1.73E+01	1.72E-03
103	210242	C	T		Downstream gene			1.93E+01	6.74E-04	1.82E+01	1.14E-03
120	27573	A	G	GSONAT00011726001	Synonymous	p.Gly48Gly	None predicted	1.82E+01	1.13E-03	1.71E+01	1.81E-03
126	151167	G	A	GSONAT00011915001	Synonymous	p.Thr87Thr	None predicted	1.98E+01	5.58E-04	1.74E+01	1.65E-03
126	151204	A	G	GSONAT00011915001	Splice region&intron		None predicted	1.80E+01	1.22E-03	1.73E+01	1.68E-03
131	32856	G	T		Upstream gene			2.04E+01	4.22E-04	2.02E+01	4.46E-04
140	42203	A	G	GSONAT00012258001	3 prime UTR		Innexin	1.80E+01	1.22E-03	1.83E+01	1.08E-03
169	20585	G	A	GSONAT00012792001	3 prime UTR		Glutamic rich SH3 binding domain	1.82E+01	1.11E-03	1.80E+01	1.26E-03
175	29781	T	G		Upstream gene			1.71E+01	1.82E-03	1.70E+01	1.97E-03
196	842	T	A	GSONAT00013101001	Missense	p.Lys112Met	None predicted	2.37E+01	9.37E-05	2.23E+01	1.72E-04
212	24957	G	A	GSONAT00015370001	NA			1.98E+01	5.58E-04	1.85E+01	1.01E-03
212	45266	A	T	GSONAT00013238001	Synonymous	p.Ala493Ala	Peroxidase	1.84E+01	1.02E-03	1.79E+01	1.31E-03

262	14087	C	T	GSONAT00013467001	Synonymous	p.Ser60Ser	None predicted	2.22E+01	1.86E-04	2.13E+01	2.82E-04
360	660	G	A	GSONAT00015450001	NA		Uncharacterized protein	1.76E+01	1.50E-03	1.62E+01	2.81E-03
408	5071	G	A	GSONAT00015430001	Intron		None predicted	1.75E+01	1.54E-03	1.67E+01	2.18E-03
408	6331	C	A	GSONAT00015430001	Intron		None predicted	1.94E+01	6.47E-04	1.79E+01	1.28E-03
541	2367	C	T	GSONAT00013822001	Missense	p.Glu1091Lys	LNR domain	1.81E+01	1.16E-03	1.67E+01	2.21E-03
556	3426	C	T	GSONAT00015380001	NA		LNR domain/Kelch domain	2.18E+01	2.15E-04	2.05E+01	4.06E-04
556	3591	C	T	GSONAT00015380001	NA		LNR domain/Kelch domain	2.04E+01	4.22E-04	1.89E+01	8.35E-04
1090	1757	C	G	GSONAT00014235001	Intron		None predicted	1.95E+01	6.25E-04	1.82E+01	1.12E-03
1239	1585	C	T		Intergenic			2.37E+01	9.37E-05	2.23E+01	1.72E-04
1239	606	T	C	GSONAT00015400001	NA		LNR domain/metallopeptidase	2.37E+01	9.37E-05	2.23E+01	1.72E-04
1365	1534	C	G	GSONAT00014370001	Missense	p.Glu699Asp	Laminin subunit	1.85E+01	9.71E-04	1.82E+01	1.15E-03
1365	2873	C	A	GSONAT00014370001	Missense	p.Asp337Tyr	Laminin subunit	1.84E+01	1.01E-03	1.68E+01	2.07E-03
1429	1120	A	G		Upstream gene			1.79E+01	1.28E-03	1.63E+01	2.62E-03
1604	1098	G	A	GSONAT00014466001	Synonymous	p.Ile487Ile	FAD/NAD(P)-binding domain	1.79E+01	1.31E-03	1.58E+01	3.28E-03
1807	2980	G	C	GSONAT00015410001	NA		LNR domain	2.37E+01	9.37E-05	2.23E+01	1.72E-04
1819	1461	C	G	GSONAT00014573001	Missense	p.Pro171Ala	Kelch-type beta propeller domain	1.85E+01	9.80E-04	1.89E+01	8.07E-04
1819	2618	C	T	GSONAT00014573001	Splice region&intron		Kelch-type beta propeller domain	2.15E+01	2.48E-04	2.07E+01	3.66E-04
1819	2868	T	C	GSONAT00014573001	Missense	p.Tyr483His	Kelch-type beta propeller domain	1.89E+01	8.14E-04	1.83E+01	1.08E-03
1886	3092	A	T		Downstream gene			1.73E+01	1.72E-03	1.70E+01	1.97E-03
2017	1371	G	A		Intergenic			2.01E+01	4.84E-04	2.08E+01	3.44E-04
2017	1736	G	T		Intergenic			1.74E+01	1.63E-03	1.71E+01	1.82E-03
2023	2729	A	G		Upstream gene			2.37E+01	9.37E-05	2.23E+01	1.72E-04
2066	3045	T	A		Intergenic			1.86E+01	9.42E-04	1.76E+01	1.50E-03
2085	1564	T	C	GSONAT00014698001	Missense	p.His392Arg	LNR domain	2.08E+01	3.43E-04	2.00E+01	5.02E-04
2085	1603	C	G	GSONAT00014698001	Missense	p.Trp379Ser	LNR domain	1.92E+01	7.29E-04	1.80E+01	1.23E-03
2085	2429	T	G	GSONAT00014698001	Missense	p.Thr104Pro	LNR domain	2.25E+01	1.59E-04	2.04E+01	4.11E-04
2487	2406	T	G	GSONAT00015390001	NA		Kelch-type beta propeller domain	2.15E+01	2.48E-04	2.08E+01	3.44E-04
3137	1880	G	T	GSONAT00015041001	Missense	p.Phe109Leu	Gamma-glutamyltranspeptidase	2.13E+01	2.78E-04	1.94E+01	6.46E-04
Scaffold	Position	Ref	Alt	Gene model	Variant	AA modification	Annotation	Lk.b	pval.Lk.b	Lk.d	pval.Lk.d
3137	1912	T	C	GSONAT00015041001	Missense	p.Lys99Glu	Gamma-glutamyltranspeptidase	2.19E+01	2.09E-04	1.98E+01	5.45E-04
3250	232	C	T		Intergenic			1.82E+01	1.15E-03	1.70E+01	1.96E-03
3397	1439	T	C		Intergenic			1.87E+01	8.96E-04	1.77E+01	1.40E-03
3651	2020	G	A		Intergenic			1.94E+01	6.47E-04	1.72E+01	1.81E-03

Supplementary Notes S10: MGVs in the *Tara* Oceans sampling stations. Samples for which the sequencing was not performed are marked by '-'.

			Fraction size			
Stations	**Latitude**	**Longitude**	**0.8-5 µm**	**5-20 µm**	**20-180 µm**	**180-2 000 µm**
4	36.0552	-6.56275	417 412	25 713	-	339 857
7	37.04875	1.9402	10 108	-	381 729	-
9	39.12085	5.85945	515 141	-	-	388 253
11	41.1686	2.7996	-	28 744	301 314	553 424
18	35.7491	14.29475	413 215	12 892	472 164	-
20	34.44785	14.9261	-	350 947	-	561 809
22	39.83935	17.41535	495 999	231 644	626 211	575 398
23	42.1956	17.71695	615 319	-	675 132	359 567
25	39.37435	19.39945	330 787	269 951	553 585	480 709
26	38.4493	20.18135	-	-	592 680	-
30	33.9191	32.86955	-	202 255	800 919	509 937
66	-34.92175	17.97445	355 128	171 295	667 244	645 468
67	-32.2176	17.70515	475 894	373 570	862 323	657 083
68	-31.0379	4.66455	1 184 712	212 138	446 406	609 082
70	-20.40335	-3.1884	824 094	191 715	753 658	435 722
72	-8.81015	-17.91365	-	360 995	591 716	434 908
76	-20.97135	-35.25735	-	16 664	707 466	529 141
78	-30.1626	-43.2865	-	196 417	589 175	885 883
80	-40.65025	-52.1683	508 566	-	327 369	464 035
81	-44.5366	-52.43165	883 668	-	812 376	-
82	-47.18845	-58.2792	975 148	11 381	652 254	346 936
83	-54.3766	-65.07765	743 188	391 588	-	431 141
84	-60.29465	-60.5786	179 031	353 092	643 417	519 925
85	-62.0874	-49.43105	406 706	-	576 267	255 892
142	25.546	-88.42815	512 077	-	722 244	368 175
143	29.6871	-79.60905	-	291 332	570 557	344 135
144	36.36745	-72.86875	309 181	93 936	717 114	371 912
145	39.2239	-71.0352	512 608	164 452	484 753	412 953
146	34.75265	-71.25665	434 823	16 041	598 415	433 382
147	33.01575	-66.53725	1 223 319	266 103	337 447	257 342
148	31.7657	-64.17895	-	127 807	603 575	576 576
149	34.1129	-49.92245	-	132 183	246 440	348 584
150	35.89685	-37.238	813 156	9 281	414 525	537 250
151	36.15525	-28.9914	1 100 155	28 218	297 788	331 450
152	43.6849	-16.84495	823 671	306 020	405 186	559 005